ONLINE METHODS

Standard MLM. A standard MLM for GWAS can be written by extending the notation of Henderson[15] as follows:

$$\mathbf{y} = W\mathbf{v} + X\boldsymbol{\beta} + Z\mathbf{u} + \mathbf{e} \quad (1)$$

where $\mathbf{y}$ is a vector of a phenotype; $\boldsymbol{\nu}$ and $\boldsymbol{\beta}$ are unknown fixed effects representing marker effects and non-marker effects, respectively; and $\mathbf{u}$ is a vector of size n (number of individuals) for unknown random polygenic effects having a distribution with mean of zero and covariance matrix of $G = 2K\sigma_a^2$, where K is the kinship (co-ancestry) matrix with element k_{ij} ($i,j = 1,2,\ldots,n$) calculated from either a set of genetic markers or pedigrees and σ_a^2 is an unknown genetic variance. W, X and Z are the incidence matrices for ν, β and $\mathbf{u}$, respectively, and $\mathbf{e}$ is a vector of random residual effects that are normally distributed with zero mean and covariance $\mathbf{R} = I\sigma_e^2$, where I is the identity matrix and σ_e^2 is the unknown residual variance. The null hypothesis for the association test that is $\nu = 0$ and the alternative hypothesis is that $\nu \neq 0$. The test of the null hypothesis can be performed by either an F test or χ^2 test after the maximization of the following likelihood:

$$L(\mathbf{y}|\mathbf{v},\beta,\mathbf{u},\sigma_a^2,\sigma_e^2) \quad (2)$$

Compression. The form of the compressed MLM is the same as equation (1). The difference in content is that individuals in u are replaced by their corresponding groups, and kinship among individuals (K) is replaced by the kinship among groups ($\bar{k}$), which is defined as $\bar{k} = \{\bar{k}_{ij}\}$, where $i,j = 1$ to s, and where

$$\bar{k}_{ij} = \underset{h\in i,\, t\in j}{\text{average}}(k_{ht}) \quad (3)$$

Under the compressed MLM, the likelihood (L) is as follows:

$$L(y|\mathbf{v},\beta,\mathbf{u},\sigma_a^2,\sigma_e^2,C) \quad (4)$$

where $\mathbf{C}$ is the clustering results after using a clustering algorithm with s groups (where $s = 1,2,\ldots,n$).

P3D. The first step of P3D is to determine population parameters, including genetic variance (σ_a^2), residual variance (σ_e^2) and clustering result (C), by maximizing the following likelihood:

$$L(y|\beta,\mathbf{u},\sigma_a^2,\sigma_e^2,C) \quad (5)$$

Then, with the population parameters fixed as empirical Bayesian priors[23], the non-population parameters (v, β and u) are optimized for each marker by maximizing the following likelihood:

$$L(y|\mathbf{v}\,\beta,\mathbf{u},\hat{\sigma}_a^2,\hat{\sigma}_e^2,\hat{C}) \quad (6)$$

Equation (6) is maximized by solving equation (1) only once (no iteration) while holding those population parameters constant.

Observed data. We examined three genetic association datasets from human, dog and maize. Each dataset contained phenotype data and a set of genetic markers.

The human dataset was collected from 1,315 adult individuals (specifically, European Americans over 17 years old) who participated in the Genetics of Lipid Lowering Drugs and Diet Network (GOLDN) study[28]. There were 637 genetic markers (388 microsatellite, or simple sequence repeat, markers and 259 SNP markers) scored on these individuals. All multiallelic simple sequence repeat markers were converted into biallelic markers by collapsing alleles into two states: the major allele and all other alleles. Measured phenotypes included height, physical activity, and serum triglyceride and cholesterol levels. Age and sex were also recorded for each individual. A prior study[28] found no significant population structure in this population and no statistically significant association between height and the genetic markers.

The dog dataset was based on 292 dogs from two breeds (Labrador retriever and greyhound) and their crosses (F_1, F_2 and two backcrosses). Hip dysplasia was indicated by Norberg angle measured on both the left and right sides. The lowest hip score (the minimum between the left and right measurements) was used in the analysis[29]. All dogs were genotyped with 23,500 SNPs at genome-wide coverage, of which 1,000 SNPs were randomly sampled for this study.

The maize dataset was composed of phenotype (flowering time scored as days to pollination), genotype (553 SNPs) and population structure (Q matrix) in 277 inbred lines[5]. No statistically significant association was found between the genetic markers and flowering time. This dataset is downloadable as a tutorial dataset of the TASSEL software package[27].

Simulation schemes. Two schemes were employed to simulate phenotypes each for the examination of compressed MLM and P3D. In both schemes, we used SNP marker data from the human, dog and maize datasets. Also, in each scheme, the population structure effect and impact of kinship were retained.

Scheme 1 was to add additional QTN effects to an observed phenotype[5]. This scheme was used to evaluate the compressed MLM approach on phenotypes with the original genetic architecture being retained. The added QTN effect contributed to only a small proportion of variation in that phenotype (0.03%–6.00%).

The QTN effect was represented in the unit of phenotypic standard deviation (k). The percentage of the total variation explained by the QTN (π) is a function of k and sample frequency (f) of the polymorphism at the QTN, defined as $1/(1+1/f\,(1-f)k^2)$[30]. Larger effects (a maximum of $k = 0.5$) were added for the dog and maize datasets, in which the sample sizes were smaller. Smaller effects (a maximum of $k = 0.2$) were added to the human dataset, which had a larger sample size sufficient to allow a small QTN effect to be detected. For a QTN with the largest effect ($k = 0.5$), the percentage of the total variation explained reaches a maximum value of 5.88% when $f = 0.5$. To facilitate comparison between datasets, we listed π at the $f = 0.3$. The genetic effect was assigned to all SNPs, one at a time, to produce replicates across all SNPs.

Scheme 2 was to simulate a phenotype with every known element, including the contribution of population structure, genetic effects (additive, dominance and epistatic) and residual effect. We used this scheme to examine whether P3D could work across traits with different genetic architecture. The general equation to simulate a phenotype (y) is as follows:

$$y = \text{population structure} + \text{additive} + \text{dominance} + \text{epistatic} + \text{residual} \quad (7)$$

where 'population structure' was based on the first five principal components, which were derived from all the genetic markers. The population structure explained 1% of the total phenotypic variation for humans, 25% for dogs and 25% for maize. 'Additive' is the sum of all additive effects for a known number of causal QTNs (5 or 20). The distribution of these QTN effects followed a geometric series[31]. The effect of the i^{th} QTN was set as a^i, where $a = 0.92$. The proportion of the additive effect was defined by the narrow-sense heritability (h^2), which is the proportion of additive variance over the total variance (sum of additive and non-additive variances). Non-additive variance (dominance, epistatic and residual) was set to $V_a(1-h^2)/h^2$, where V_a is the additive genetic variance calculated among the total additive genetic effects across QTNs for each individual. Two levels of heritability were examined ($h^2 = 0.25$ or 0.5). 'Dominance' is the sum of dominance effects from all QTNs with a dominance effect of da^i for heterozygotes at the i^{th} QTN, where d is the degree of dominance ($d = 0$, 0.25, 0.5 or 1). 'Epistatic' is the sum of pairwise interaction effects among all QTNs. The magnitude of the epistatic effect is indicated by the proportion of total variance explained by the epistatic effect (proportion of variance of 0, 0.05, 0.1 and 0.2). The 'residual' effect follows a normal distribution and has a variance to satisfy the contributions from additive, dominance and epistatic effects at the designated level. Simulations of the phenotypes were repeated 1,000 times. The non-causal SNPs were randomly sampled q times for each replicate, where q was set to the same number of QTN in each scenario ($q = 5$ or 20).

Statistical analysis. Proc mixed in SAS[26] was used to solve the MLM with variance components estimated by the restricted maximum likelihood algorithm. Model fit was examined with three criteria: negative log likelihood, adjusted Akaike information criterion and Bayesian information content.

For the analysis of the human dataset, the fixed effects were sex, age and the quadratic term of age in the evaluation of the observed phenotypes

and phenotypes simulated under scheme 1. Similarly, breed (or fraction of Labrador retriever, for the crosses with greyhound) was the fixed effect in the analysis of the dog dataset, and population structure was the fixed effect in the analysis of the maize dataset. The first five principal components[6] derived from all genetic markers were fit as fixed effects for the phenotypes simulated under scheme 2.

Individuals or their corresponding groups were fit as a random effect. The kinship among individuals was estimated from the genetic markers by the approach of Loiselle *et al.*[32]. The individuals in each dataset were grouped based on their kinship by using proc cluster in SAS[26]. The genotypic effect of each genetic marker was fit as a fixed effect, one marker at a time. The association tests on the markers' genotypes were performed by conducting F tests.

URLs. Compression and P3D were implemented in SAS (**Supplementary Note**) and TASSEL[27] software package. The SAS code, standalone TASSEL program and demonstration date are available at http://www.maizegenetics.net/.

28. Lai, C.Q. *et al.* Fenofibrate effect on triglyceride and postprandial response of apolipoprotein A5 variants: the GOLDN study. *Arterioscler. Thromb. Vasc. Biol.* **27**, 1417–1425 (2007).
29. Zhang, Z. *et al.* Estimation of heritabilities, genetic correlations, and breeding values of four traits collectively defining hip dysplasia in dogs. *Am. J. Vet. Res.* **70**, 483–492 (2009).
30. Long, A.D. & Langley, C.H. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res.* **9**, 720–731 (1999).
31. Lande, R. & Thompson, R. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* **124**, 743–756 (1990).
32. Loiselle, B.A., Sork, V.L., Nason, J. & Graham, C. Spatial genetic-structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *Am. J. Bot.* **82**, 1420–1425 (1995).

doi:10.1038/ng.546